Air Quality Assessment and Pollution Forecasting using Recurrent Artificial Neural Networks in Metropolitan Lima-Peru

Chardin Hoyos Cordova
Universidad Peruana Unión
Manuel Niño Lopez Portocarrero
Universidad Peruana Unión
Rodrigo Salas
Universidad de Valparaíso
Romina Torres
Universidad Andres Bello
Paulo Canas Rodrigues
Federal University of Bahia
Javier Linkolk López-Gonzales (javierlinkolk@gmail.com)
Universidad de Valparaíso

Research Article

Keywords: Long-Short Term Memory networks (LSTM), hold-out (HO), blocked-nested cross-validation (BNCV), World Health Organization (WHO), Environmental Protection Agency (EPA)

Posted Date: September 9th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-869832/v1

Version of Record: A version of this preprint was published at Scientific Reports on December 1st, 2021.
See the published version at https://doi.org/10.1038/s41598-021-03650-9.

Air Quality Assessment and Pollution Forecasting using Recurrent Artificial Neural Networks in Metropolitan Lima-Peru

Chardin Hoyos Cordova[1], Manuel Niño Lopez Portocarrero[1], Rodrigo Salas[2], Romina Torres[3], Paulo Canas Rodrigues[4], and Javier Linkolk López-Gonzales[1,5,*]

[1]Facultad de Ingeniería y Arquitectura, Universidad Peruana Unión, Lima 15, Perú
[2]Escuela de Ingeniería C. Biomédica, Universidad de Valparaíso, Valparaíso, Chile
[3]Engineering Faculty, Universidad Andres Bello, Viña del Mar, Chile
[4]Department of Statistics, Federal University of Bahia, Salvador, Brazil
[5]Instituto de Estadística, Universidad de Valparaíso, Valparaíso, Chile
[*]javierlinkolk@gmail.com

ABSTRACT

The prediction of air pollution is of great importance in highly populated areas because it has a direct impact on both the management of the city's economic activity and the health of its inhabitants. In this work, the spatio-temporal behavior of air quality in Metropolitan Lima was evaluated and predicted using the recurrent artificial neural network known as Long-Short Term Memory networks (LSTM). The LSTM was implemented for the hourly prediction of PM_{10} based on the past values of this pollutant and three meteorological variables obtained from five monitoring stations. The model was evaluated under two validation schemes: the hold-out (HO) and the blocked-nested cross-validation (BNCV). The simulation results show that periods of low PM_{10} concentration are predicted with high precision. Whereas, for periods of high contamination, the LSTM network with BNCV has better predictability performance. In conclusion, recurrent artificial neural networks with BNCV adapt more precisely to critical pollution episodes and have better performance to forecast this type of environmental data, and can also be extrapolated to other pollutants.

Introduction

The World Health Organization (WHO) reports that air pollution causes 4.2 million premature deaths per year in cities and rural areas around the world[1]. The US Environmental Protection Agency[2] mentions that one of the pollutants with the most significant negative impact on public health is particulate material with a diameter of less than μm (PM_{10}) because it can easily access the respiratory tract, causing severe damage to health. For their part, Valdivia and Pacsi[3] report that Metropolitan Lima is vulnerable to high concentrations of PM_{10}, due to its accelerated industrial and economic growth, in addition to its large population, as it is home to 29% of the total Peruvian population[4].

To mitigate the damage caused by PM_{10} to public health, the WHO established concentration thresholds suitable to achieve a minimum adverse effect on health[5]. In various countries, several laws were issued to regulate PM_{10} concentrations and air quality in general[6], as established in Peru by the Ministry of the Environment[7] and in, e.g., the United States by the Environmental Protection Agency (EPA)[8].

In recent years, various forecasting methodologies have been adapted and developed to understand how pollutants behave in the air at the molecular level, simulating diffusion and dispersion patterns based on the size and type of the molecule. However, the results of the prediction tend to achieve somehow low precision[9,10]. Examples of such models are the Community Multiscale Air Quality Model and Weather Research and Forecasting model coupled with Chemistry developed in Chen et al.[11] and Saide et al.[12] respectively, which are used to forecast air quality in urban areas. On the other hand, some methods tend to be more appropriate to model and forecast air quality because they use statistical modeling techniques, such as Artificial Neural Networks (ANNs). These kind of models have been widely used to forecast time series in general and applied to environmental data such as particulate matter in different countries[13,14], in particular.

Several studies have been focusing on applying recurrent neural networks to forecast air quality in large cities. For instance, Guarnaccia et al.[15] reported that predicting the air quality with high accuracy is a problematic issue, which is becoming increasingly important because it is a tool capable of providing complete information for helping to prevent critical pollution episodes and of reducing the human exposure to these contaminants[13,16,17]. However, there is a limited number of studies in

the context of Lima, Peru, which is one of the cities with the highest pollution in South America[18–20]. Herrera and Trinidad[21] used neural networks to predict PM_{10} in the Carabayllo district - Lima, with a good forecasting performance. In another study[22] aimed at forecasting PM_{10} three days ahead and at comparing the performance of the standard LSTM, GRU, and RNN models, concluding that all three models showed good performance for out-of-sample forecasting.

In this study, the air quality of Lima was evaluated to understand its behavior and the possible causes and factors that favor pollution. Subsequently, we applied the long short term memory (LSTM) models to forecast PM_{10} concentrations, where the model was evaluated under two validation schemes: the Hold-out (HO) and the Block Nested Cross-Validation (BNCV). Historical hourly PM_{10} data from five air quality monitoring stations were used.

Materials and Methods

In this work, we follow the Knowledge Discovery from Databases (KDD) methodology to obtain relevant information for air quality management decision-making. The main goal of the KDD is to extract implicit, previously unknown, and potentially helpful information[23] from raw data stored in databases. Therefore, the resulting models can predict, e.g., one hour ahead, the air quality and support the city's management decision-making (see Figure 1).

The KDD methodology has the following stages: (a) Phenomena Understanding; (b) Data Undersdanding; (c) Data Preparation; (d) Modeling; (e) Evaluation; and, (d) Selection/Interpretation. In the following subsections we explain each stage of the process.

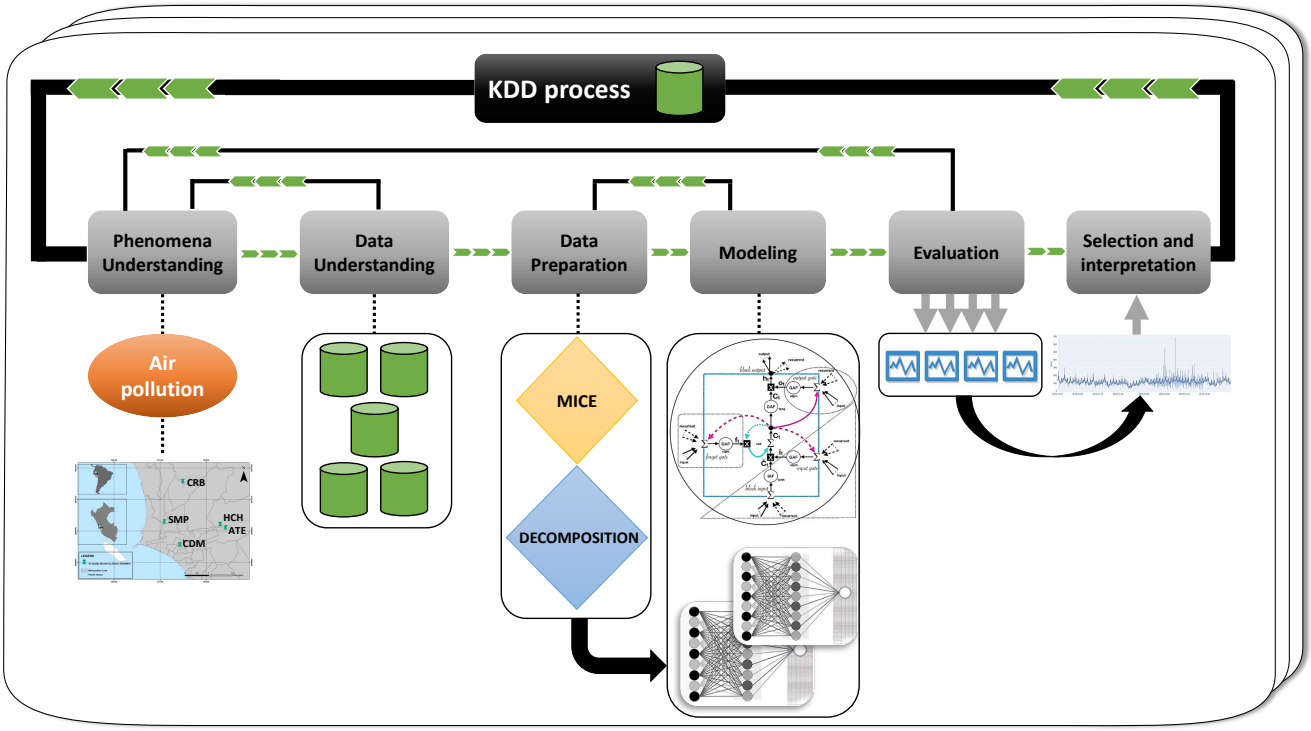


Figure 1. Knowledge Discovery from Databases (KDD) methodology used for Air Quality Assessment and Pollution Forecasting

Phenomena Understanding

At this stage, we contextualize the contamination phenomenon concerning the PM_{10} concentration in the five Lima monitoring stations. The main focus is to predict air pollution to support decision-making related to establishing pollution mitigation policies. For this, we use the LSTM as a computational statistical method for PM_{10} forecasting.

Lima is the capital of the Republic of Peru. It is located in the center of the western side of the South American continent in the 77° W and 12° S and, together with its neighbor, the constitutional province of Callao, form a populated and extensive metropolis with 10,628,470 inhabitants and an area of 2819.3 km^2[24,25].

The average relative humidity (temperature) in summer (December to March) ranges from 65% - 68% (24°C - 26°C) in the mornings, while at night the values fluctuate between 87% - 90% (18°C - 20°C). In winter (June to September), the average

daytime relative humidity ranges between 85% - 87% (18°C - 19 °C) and at night it fluctuates between 90% - 92% (18°C - 19 °C). The average annual precipitation is 10 mm. On the other hand, the average altitudes reached by the thermal inversion in summer and winter are approximately 500 and 1500 meters above sea level, respectively[26,27].

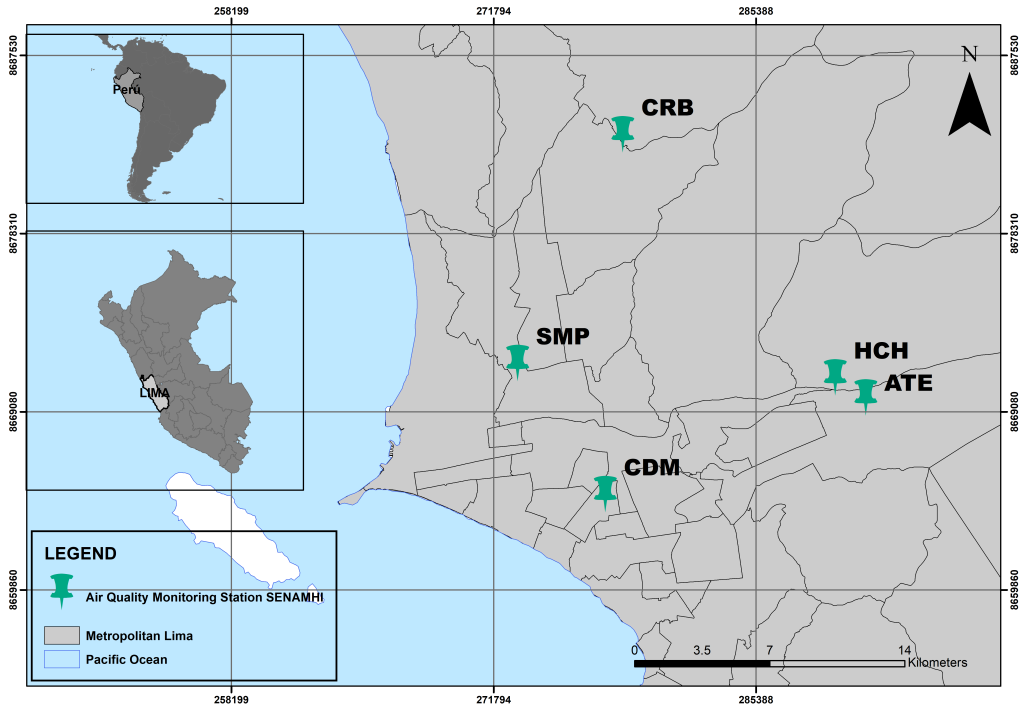


Figure 2. Location map of the study area and the Lima air quality monitoring stations: ATE, Campo de Marte (CDM), Carabayllo (CRB), Huachipa (HCH) and San Martin de Porres (SMP).

Variable	Unit of measurement
PM_{10}	$\mu g/m^3$
Temperature	°C
Relative humidity	%
Wind speed	w/s
Direction of the wind	degrees (°)

Table 1. Pollutant and weather variables used in this study

Data Understanding

Lima has ten air quality monitoring stations located in the constitutional province of Callao and the north, south, east, and center of Lima. The data used comprise hourly observations from January 1, 2017, to December 31, 2018, and includes three meteorological variables and the concentration of particulate matter PM_{10}, considered to be an agent that, when released into the environment, causes damage to ecosystems and living beings[28,29]. For this study, the hourly data, recorded at five air quality monitoring stations (see Figure 2), which are managed by the National Service of Meteorology and Hydrology (SENAMHI), was considered. Table 1 shows the considered variables and their units of measurement.

When considering environmental data, such as PM_{10} concentrations, from different locations, preliminary spatio-temporal visualization studies are of great use to better understand the behavior of the meteorological variables, the topography of the area, and pollutants[30].

Data Preparation

This stage is relevant since it precedes the modeling stage. The preparation of the data had various treatments. First, we address the problem of missing data. The treatment was performed with the MICE library. This library performs multiple imputations

using the Fully Conditional Specification[31] and requires a specification of a separate univariate imputation method for each incomplete variable. In this context, predictive mean matching, a versatile semiparametric method focusing on continuous data, was used, which allows the imputed values to match one of the observed values for each variable. The data imputation was performed for each of the five stations with percentage of missing data below 25%.

The data from the monitoring stations consist of sequence of observed values $\{x_t\}$ recorded at specific times t. In this case, the time series is collected at hourly intervals. We proceed to normalize all the records in the range [0,1] as follows:

$$X_t = \frac{x_t - \min\{x_t\}}{\max\{x_t\} - \min\{x_t\}} \tag{1}$$

Moreover, the time series is decomposed into the trend, seasonality, and the irregular components following an additive model (the cyclic component is omitted in this work):

$$X_t = Trend_t + Cyclic_t + Seasonal_t + Irregular_t \tag{2}$$

The trend component $Trend_t$ at time t reflects the long-term progression of the series that could be linear or non-linear. The seasonal component $Seasonality_t$ at time t, reflects the seasonal variation. The irregular component $Irregular_t$ (or "noise") at time t describes the random and irregular influences. In some cases, the time series has a cyclic component $Cyclic_t$ that reflects the repeated but non-periodic fluctuations. The main idea of applying this decomposition is to obtain the deterministic and non-deterministic components, where a forecasting model is obtained for the deterministic part. In this article, we have used the method implemented in Statmodels for Python[32].

Modeling

Artificial Neural Networks (ANNs) have received a great deal of attention in engineering and science. Inspired by the study of brain architecture, ANNs represent a class of non-linear models capable of learning from data[33]. The essential features of an ANN are the basic processing elements referred to as neurons or nodes, the network architecture describing the connections between nodes, and the training algorithm used to estimate values of the network parameters.

Researchers see ANNs as either highly parameterized models or semiparametric structures[33]. ANNs can be considered as hypotheses of the parametric form $h(\cdot;\mathbf{w})$, where the hypothesis h is indexed by the vector of parameters $\mathbf{w}$. The learning process consists of estimating the value of the vector of parameters $\mathbf{w}$ to adapt the learner h to perform a particular task.

Machine Learning and deep learning methods have been successfully applied for time series forecasting[34–39]. For instance, RNNs are dynamic models frequently used for processing sequences of real data step by step, predicting what comes next, and they are applied in many domains, such as the prediction of pollutants[40]. It is known that when there are long-term dependencies in the data, RNNs are challenging to train, which leads to the development of models such as the Long Short Term Memory (LSTM) that have been succesfully applied in time series forecasting[41].

The LSTM model is a type of RNN, having as its primary strength the ability to learn long-term dependencies and being a solution for long time series intervals[20,42]. In such a model, memory blocks replace the neurons in the hidden layer of the standard RNN[43]. The memory block consists of three gates that control the system's state: Input, forget, and output gates. First, the input gate determines how much information will be added to the cell. Second, the forget gate controls the information lost in the cells. Lastly, the output gate performs the function of determining the final output value based on the input and memory of the cell[44,45].

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{3}$$

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \tag{4}$$

$$\tilde{C}_t = \tanh\left(W_{\tilde{C}} \cdot [h_{t-1}, x_t] + b_{\tilde{C}}\right) \tag{5}$$

$$C_t = (f_t \cdot C_{t-1}) + \left(i_t \cdot \tilde{C}_t\right) \tag{6}$$

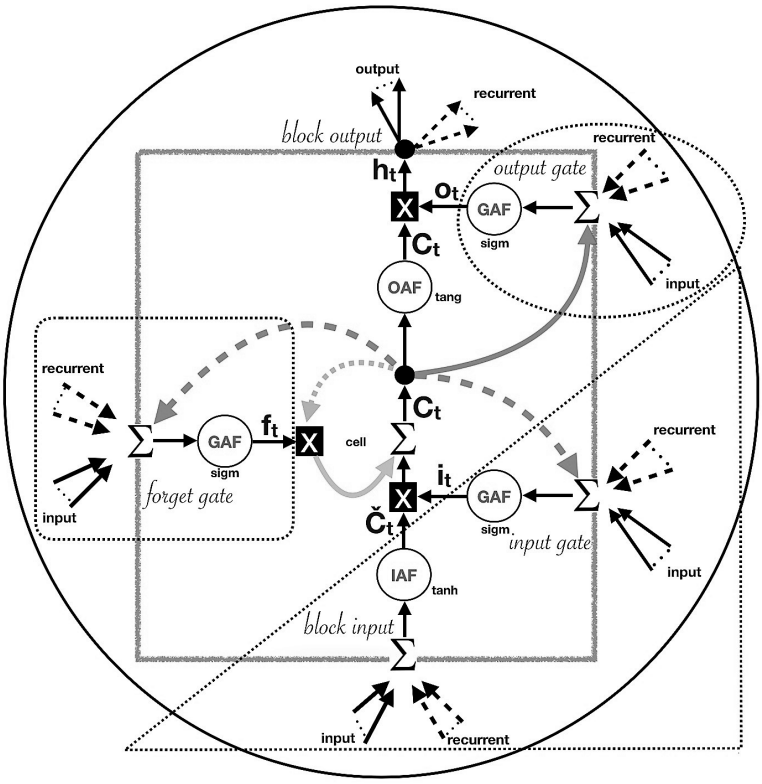


Figure 3. Model of one block of the LSTM. The block is composed of the input gate, forget gate and output gate.

$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \tag{7}$$

$$h_t = o_t * tanh(C_t) \tag{8}$$

Figure 3 shows the LSTM model block, with the output and input blocks, which consists of three gates. At each step, an LSTM maintains a hidden vector h and a memory vector o responsible for controlling status updates and outputs. The first step is to decide what information will not be considered in the status cell. This decision is made by the forget gate, which uses a hyperbolic tangent activation function (IAF). f_t represents the output of the forget gate, that can be calculated using equation (3). This gate considers the concatenation of the vectors h_{t-1} and x_t. It generates a number between 0 and 1 for each number in the state cell C_{t-1}, where W_f and b_f are the weight matrices and the bias vector parameters, respectively. Both must be learned during training and are stored in the vector f_t. If one of the values of this vector is equal to or close to zero, then the LSTM will eliminate that information. On the other hand, if it reaches values equal to or close to 1, this information will be maintained and reach the status cell. The next step is to decide what new information to store in the status cell. This is done by the input gate, linked to a sigmoid activation function (GAF), and with an output for that gate (i_t), all this is calculated by the equation (4, 5). In addition, for the input block, the hyperbolic tangent activation function (IAF) is used. First, the vectors h_{t-1} and x_t are concatenated. Being W_i and b_i, the weight matrices and the bias vector parameters, respectively, must be learned during training; all this is stored in the vector i_t called the input gate, which decides which values to update. Then a hyperbolic tangent function creates a vector of new candidate values, $\tilde{C}_t$, involving the vectors h_{t-1} and x_t. In the next step, these values are filtered by multiplying point by point both vectors to create a status cell update. The previous cell, C_{t-1} is updated to the new state of cell C_t (equation 6). In addition, the output gate, also linked with the GAF activation function and with an output of the output gate (o_t), for its calculation uses the equation (equation 7). Finally, h_t, expresses the new output of the model (equation 8). The current cell state is represented by C_t, while W is the weight of the model, and b is the bias of the model.

Predictive models aim to predict future data values. In this study, we focus on the PM_{10} concentration prediction. For this, the LSTM was used with a particular architecture, ACF-PACF. Based on the ACF (autocorrelation function) and PACF (partial

autocorrelation function), relevant lags were detected that are used in the model. The configuration of the network is associated with the information provided by the partial autocorrelation function. The information given by the ACF is $t-1$, $t-2$, $t-3$, $t-23$ and $t-24$. In addition, temperature, relative humidity, and wind speed are used with $t-4$ (4 hours ago). In this work, we have implemented a LSTM with 10 blocks, the output of each block is aggregated with a single neuron. To train the model we have used the adam optimizer and the mean squared error of the loss function, 10% of the data was used for validation to avoid over-fitting with early stopping. A maximum of 500 epochs and batch sizes of 1024 was used to fit the weights of the model.

Model Evaluation

To evaluate the forecast ability of the models, the performance metrics given below were used (see[46,47]). In what follows, we will consider: y_i, $i = 1,...,n$, are the target values; $\hat{y}_i$, $i = 1,...,n$, are the model's predictions; $\bar{y}_i$ is the mean of the target values; and n is the number of samples.

1. Mean Absolute Error: is the average of the absolute difference between the target and the predicted values.

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - \widehat{y}_i|}{n} \tag{9}$$

2. Root Mean Squared Error: is the squared root of the average of the squared errors.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{n}} \tag{10}$$

3. Symmetric Mean Absolute Percentage Error: is a measure of accuracy based on percentage of relative errors.

$$\text{sMAPE} = \frac{100\%}{n} \sum_{i=1}^{n} \frac{|y_i - \widehat{y}_i|}{|\widehat{y}_i| + |y_i|} \tag{11}$$

4. Spearman's rank correlation coefficient: is a nonparametric correlation measure between the target and the prediction. Spearman's correlation assesses monotonic relationships by using the rank of the variables.

$$S = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n\,(n^2 - 1)} \tag{12}$$

where $d_i = rg(y_i) - rg(\hat{y}_i)$ is the difference between the ranks of the targets $rg(y_i)$ and the predictions $rg(\hat{y}_i)$.

Model Selection and Interpretation

This is the final step in the KDD process and requires that the knowledge extracted from the previous step be applied to the specific domain of the PM_{10} prediction, in a visualized format. At this stage, in addition to selecting the model with the best precision in the prediction, it also drives the decision-making process based on the assessment of the air quality in Lima.

For the validation we have used two schemes: Hold-Out (HO) and Blocked Nested Cross-Validation (BNCV). On the one hand, HO has the conventional separation of the dataset in training, validation and testing sub-sets (see Fig. 4). On the other hand, the BNCV is a fixed-size window that slides; as it does so, it retrains itself with the hour it is added. That is, the model predicted the next hour, but it retrains with all the data up to the current day (see Fig. 5).

Results and Discussion

Air quality assessment in Metropolitan Lima-Peru

In this section, we report the results of the Statistical analysis of air pollution in Lima.

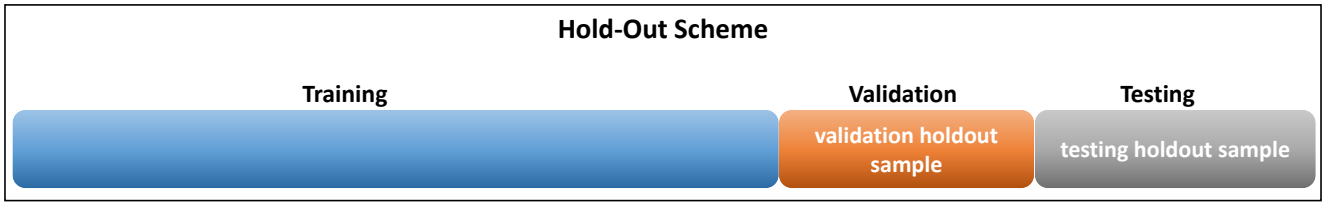


Figure 4. Hold-Out Scheme used for the validation of the models. The dataset is split into three sets: training, validation, and testing. The train set is the basis for training the model, and the test set is used to see how well the model performs in untrained PM_{10} concentrations. The division used was: 80% of the PM_{10} concentrations for training and the remaining 20% for testing.

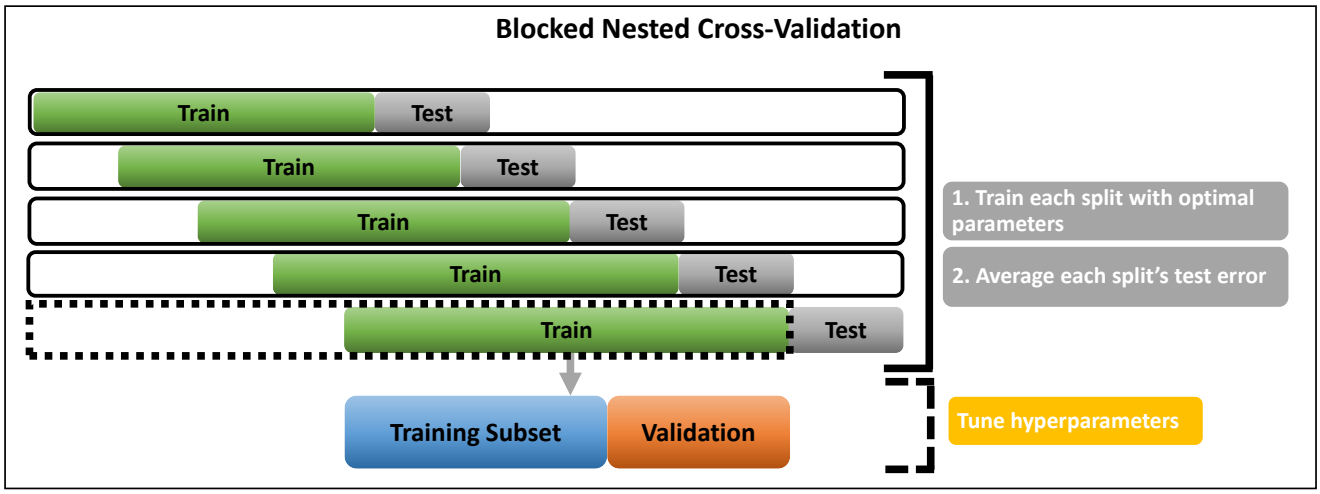


Figure 5. Blocked Nested Cross-Validation Scheme used for the validation of the models. The dataset is separated into three sets using a time-window of fixed size: training, validation, and testing. The last day is used for testing. The time-window is moved with a 1-day step for the last 60 days.

Statistical Analysis of the Concentration of PM_{10}

Table 2 shows the descriptive analysis of the data from the five monitoring stations focused in the PM_{10}, between 01-01-2017 and 31-12- 2018. Additionally, the histogram (see Figure 6) is reported to show the behavior of the pollutant in every season. In the probability distribution, it is observed that they are skewed to the right, which indicates the existence of critical episodes of contamination, being the HCH station the one with the highest incidence, with an average of 130.03 ± 91.68 $\mu g/m^3$. This value exceeds that standardized by the Peruvian norm[7], and shows relevant fluctuations and high dispersion of pollutants (8404.34 $\mu g/m^3$) that cause a high standard deviation The stations HCH and ATE register higher concentration values. The order of the stations from the lowest to the highest levels of the mean of PM_{10} is as follows: **CRB**; **CDM**; **SMP**; **ATE**; **HCH**. Similar behaviour was found in other studies[30,48]. Encalada et al.[30] carried out a study of visualization of PM_{10} concentrations in Lima using the same data, where similar behavior patterns of PM_{10} concentrations are shown in the five stations. In addition, all the stations surpass the limits established by the WHO for the maximum concentrations of PM_{10} in a year. Moreover, four of the five stations (except CRB) exceed the utmost limits of the annual arithmetic mean of PM_{10} proposed in the Quality Standards Environmental (ECA) in Peru.

Stations	Minimun	Maximun	1st Qu.	3rd Qu.	Median	Mean $\pm$ DS	Variance	Skewness	Kurtosis
CRB	5.44	488.02	31.49	58.45	198.31	48.69 ± 28.39	806.03	3.24	22.27
SMP	7.77	426.80	61.95	105.10	142.50	86.05 ± 35.73	1276.41	1.00	2.86
CDM	6.08	463.60	35.84	63.45	145.50	52.30 ± 24.61	605.54	2.30	18.25
ATE	6.41	931.00	82.90	148.00	421.90	121.56 ± 60.30	3635.75	2.08	11.07
HCH	5.21	974.00	62.10	176.50	138.40	130.03 ± 91.68	8404.34	1.53	4.89

Table 2. Descriptive statistics for the five PM_{10} monitoring stations.

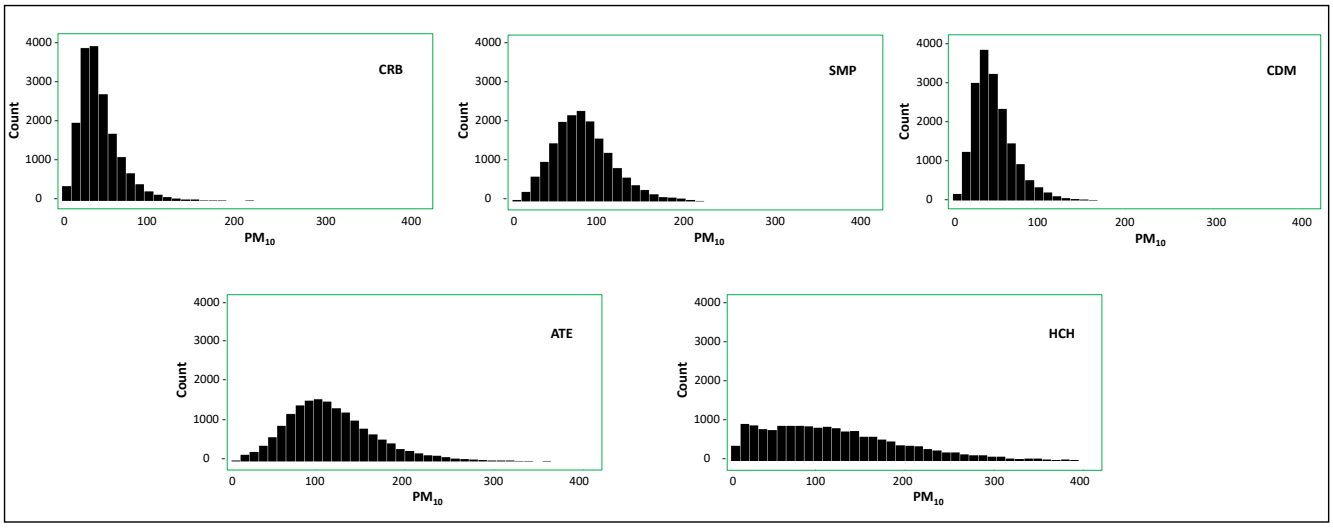


Figure 6. PM_{10} Histograms for each of the five monitoring stations, respectively CRB, SMP, CDM, ATE, and HCH.

Analysis of the Correlations with the Meteorological Variables

Regarding the correlation between PM_{10} and the meteorological variables, no significant correlation was observed, except for the station HCH, which is the area with the highest PM_{10} concentration. Figure 7 shows that there is a moderate positive correlation (0.39) between temperature and PM_{10} and a moderate negative correlation (-0.38) between relative humidity and PM_{10}. This is due to the meteorological patterns that occur in the study area. According to Silva et al.[49] between the years 1992 and 2014, the base of thermal inversions in Lima ranged between 0.6 and 0.9 kilometers from June to November, and between 0.1 and 0.6 kilometers from December to May, having a minimum average of 0.13 kilometers in March, which coincides with the season that presents critical episodes of PM_{10} concentrations.

The thermal inversion in the summer months reduces the dispersion of atmospheric pollutants because the density of the stratiform clouds decreases. Consequently, solar radiation leads to an increase in temperature and to a reduction in relative humidity. The latter results in a turbulent process causing the resuspension of coarse particles as PM_{10}[24]. High temperatures increase the photochemical activity that causes the decomposition of matter and, consequently, the increase of PM_{10}[50]. On the other hand, stratiform cloudiness increases in winter, as does relative humidity, that accompanied by drizzles in that season, help to significantly decrease the temperature and PM_{10} concentrations due to wet deposition typical of the season[27]. The above explains the high negative correlation observed between temperature and relative humidity in the five monitoring stations (see Figure 7), which is a normal phenomenon because the relative humidity directly depends on temperature and pressure to determine the capacity of the air in the intake of water vapor[51]. For this reason, the higher the temperature, the lower the relative humidity, as shown in Figure 8.

Influence of wind direction and speed on PM_{10} concentrations

The stations located in the highest area (eastern part) of the city have the highest concentration of PM_{10}. Contrary to the above, the stations located in the lowest area have a lower concentration of PM_{10}. This trend is due to the entry direction of persistent local winds from the coast to the south-southwest, which causes that pollutants such as PM_{10} are transferred to the northeast and east areas of the city, making them in critical places of contamination by particulate matter[27,30].

Although there is no significant correlation between wind speed and PM_{10}, this parameter has meteorological influence on the dispersion, resuspension, and horizontal transport of pollutants, provided that there are strong air currents (winds)[51,52], which is not the case of the present study because the highest frequencies of wind speeds are between 0 – 3.10 m/s[30].

Critical episodes of PM_{10} contamination at the HCH station

The station with the highest average PM_{10} concentration between 2017 and 2018 is HCH (see Table 2). This area has the characteristic of high vehicular traffic compared to the rest of the stations considered. The Ramiro Prialé highway, that crosses HCH and that is the most used to access the central road, connects the center and the east of the Peruvian territory, turning it into a place of high traffic congestion. Moreover, 2,462,321 vehicles were circulating in Lima[53] in 2017, and according to the National Institute of Statistics and Informatics (INEI), the vehicle fleet in Peru grew by 4.4% between 2017 and 2018[54]. The aforementioned explains the influence of high traffic vehicles in critical pollution episodes in HCH, which according to what is referred by Srishti et al.[55], the traffic caused from vehicles contributes to about 21% of PM_{10} of the pollution. In addition, it is

Metrics	MS-ATE		MS-CDM		MS-CRB		MS-HCH		MS-SMP	
	HO	BNCV	HO	BNCV	HO	BNCV	HO	BNCV	HO	BNCV
MAE	27.773	27.856	9.673	9.736	6.616	6.467	43.809	45.366	10.646	10.016
RMSE	47.561	47.229	13.867	14.146	10.810	11.083	66.406	66.706	16.380	17.047
SMAPE	24.025	24.444	19.443	19.581	17.494	17.114	34.493	36.710	14.944	13.965
Spearman r	0.513	0.510	0.665	0.658	0.754	0.763	0.640	0.621	0.810	0.816

Table 3. Results for PM_{10} prediction models for the five monitoring stations with different setups for the LSTM (without decomposition)

associated with the wear of tires and brakes[53].

Another particular feature of HCH in comparison to the other stations is the large number of tracks dilapidated, unpaved roads and the frequent inadequate disposal of land clearing on public roads by the population. These conditions generate a significant increase in dust, the main component of particulate matter, contributing to 54% of air pollution. The soil dust has a more significant impact in seasons or areas with little rainfall[55–57]. Furthermore, Lima is considered a city where it seldom rains and that only little drizzles or wet haze breakouts from cloud-type clouds nimbostratus[58].

In the surrounding area of HCH, there is also high industrial activity. The industrialization is directly associated with the increased generation of PM_{10}[58]. Concepción and Rodríguez[59] note that both the industrial activity and the vehicle fleet are the leading causes of the generation of high concentrations of PM_{10} in Lima, where the primary industries are brick kilns and non-metallic ore extraction. Moreover, was evidenced that the HCH brick industries do not have the appropriate technology to mitigate air pollution and that in all their processes, high emission of particulate matter, from the movement of land to the burning of tires, plastics, or firewood in the ovens[60]. Added to all this, it is the lack of green areas in HCH, which facilitates the resuspension of PM_{10}.

Exploratory analysis on a daily and monthly scale

The predominant time scale in the concentration of PM_{10} was evaluated in two episodes (see Figure 9). That between 07:00 and 11:00 in the morning, followed by the one between 17:00 and 22:00 at night. Similar results were found by Sánchez and Ordoñez[26], where the air quality of Lima was evaluated in 2015. From the above, it can be inferred that the levels of environmental pollution referring to PM_{10}, find the highest peaks in the evening (153.9991 and 151.9256 $\mu g/m^3$), while the lowest peaks are between 03:00 and 04:00 a.m. each day, which coincides with the results reported for the station HCH. As mentioned by Valdivia and Pacsi[3], this is related to the reduction in emissions from mobile sources that are own of the dawn.

The behavior of concentration levels of contamination varies depending on the month. In each monitoring station we can see two main peaks (see Figure 9). The first corresponds to February, March, and April, which report the highest contamination in the first semester of the year. In this period, it is the beginning of classes for schoolchildren that intensifies vehicle activity. Being in the end of the summer and the beginning of the autumn, it is a period associated with the time at which the thermal inversion occurs, which favors the generation of high peaks of PM_{10} contamination[49]. The second peak involves the winter season and the beginning of spring, highlighting mainly October as part of the second semester of the year. Similar results were found by Encalada et al.[30].

In these time windows, the stations with the highest critical episodes were HCH and ATE, while CRB had the lowest PM_{10} concentrations. In addition, from the emissions of high traffic vehicular and fixed sources of pollution, the meteorological and topographic conditions of the study area cause the high emission of PM_{10} in the air, exceeding the proposed standards in all cases by WHO.

Air Pollution Forecasting Results

Two alternatives were considered to obtain out-of-sample forecasts (see Figure 10). On the one hand, the LSTM was adjusted with the training set only once for the Hold-out scheme, and the resulting model was used to forecast one hour ahead for the last 60 days of data. On the other hand, the LSTM was trained several times with a fixed sliding window for the blocked nested cross-validation, where the model was updated for each subsequent day belonging to the test set and the following days (24 samples) were used for the test set.

The purpose of incorporating exogenous variables in this study is to improve the precision of the forecast. The exogenous variables are crucial to improve the efficiency of predictions by identifying the important meteorological covariates that affect PM_{10}, such as temperature, relative humidity, and wind speed[61].

The results show that periods of low PM_{10} concentration are predicted with very high precision. While for periods of high contamination, the model's accuracy is diminished, although in any case, it has a good degree of predictability.

Metrics	MS-ATE		MS-CDM		MS-CRB		MS-HCH		MS-SMP	
	HO	BNCV	HO	BNCV	HO	BNCV	HO	BNCV	HO	BNCV
MAE	**26.806**	**26.833**	10.723	10.745	8.551	8.495	49.751	49.371	14.240	14.138
RMSE	**41.132**	**41.183**	14.505	14.526	12.421	12.414	64.122	64.225	19.581	19.501
SMAPE	**23.470**	**23.516**	21.619	21.661	22.166	21.974	40.503	40.234	19.632	19.520
Spearman r	**0.559**	**0.557**	0.580	0.575	0.588	0.591	0.616	0.611	0.636	0.640

Table 4. Results for PM_{10} prediction models for the five monitoring stations with different setups for the LSTM (with decomposition)

Table 3 and Table 4 show the performance of the LSTM Hold-out (HO) and the LSTM blocked nested cross-validation (BNCV) schemes, without and with decomposition, respectively. RMSE shows an higher value due to extreme values in PM_{10}, being MAE not influenced by this type of values. The training data used for each station is the same in data field, acquisition time, and preprocessing. However, the experimental results show that the average RMSE values in CDM, CRB and SMP are not significantly different, but the RMSE between ATE and HCH stations shows differences. In industrial areas and heavy traffic stations (ATE and HCH), superlative changes in PM_{10} also result in higher values for the RMSE. It has been determined that the data field we are using prevents the LSTM from accurately learning the peak pollution values from traffic and industrial emissions.

The comparison in Table 3 indicates that ATE has similar prediction accuracy compared with that of HCH from the viewpoint of both the RMSE and SMAPE values. The best performance of the LSTM model with its two configurations, was seen in the SMP station. In parallel, in Table 4, similar results are observed, however, the LSTM model, with both HO and BNCV, with decomposition, presented better performance in the ATE station when compared with Table 3. This results from the fact that ATE and HCH show a greater number of critical episodes, and the model manages to adapt with some precision to the concentrations of PM_{10}. Otherwise, the model shows good performance against standard episodes, those that are within the allowed limits.

The LSTM can use the data from the previous time period to accurately forecast the value of the PM_{10} concentration in a short period of time ahead. It can learn the PM_{10} concentration trends accurately. When a station is subject to unpredictable external sources of pollution or due to short-term changes in climate and landforms (ATE and HCH), it cannot predict with accuracy. In this sense, the LSTM BNCV was able to better adapt to data from the monitoring stations that present episodes of extreme values.

Conclusions

This study addressed the problem of forecasting PM_{10} concentration on an hourly scale based on air quality indicators from five monitoring stations in Lima, Peru. A comparative study with four different configurations was made, based on the LSTM neural network: Hold-out and Blocked Nested Cross-Validation without decomposition, and Hold-out and Blocked Nested Cross-Validation with decomposition.

Operating a high-performance model in air quality forecasting in large cities, such as Lima, is one of the most critical health protection and prevention tools. Deep learning neural networks (LSTM) are recommended methodologies for designing public policies that prioritize improving air quality conditions to develop more sustainable cities.

The different configurations of the LSTM respond to the forecast of PM_{10} events by selecting the relevant meteorological variables. Precisely, the most important property of the LSTM is that through its memory units, they remember synoptic patterns over time, which is beneficial when forecasting PM_{10}. In this sense, LSTM BNCV was able to better adapt to data from the monitoring stations that present episodes of extreme values, mainly at ATE and HCH stations. On the other hand, in the stations with episodes that are within the allowed limits, the model with both configurations showed good performance.

The results show that the PM_{10} concentration prediction achieves better results with artificial intelligence methods since they are suitable for this type of approach. However, it is proposed to conduct this type of study with other cross-validation methods and hybrid and ensemble methods, which could give greater precision in the prediction. This will help in decision-making regarding air pollution mitigation and strategies, not only in Lima but also in other cities in the country and abroad. In this sense, this study of PM_{10} could be extrapolated to other pollutants, both in a national and international level.

As future work, we expect to apply other variants of deep learning models that include incremental learning[62], as well as to introduce self-identification techniques for the model identification[38,63] .

References

1. Organización Mundial de la Salud. Calidad del aire y salud. https://www.who.int/es/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health (2018).
2. Agency, U. E. P. Integrated science assessment (isa) for particulate matter (2009). EPA/600/R-08/139F.
3. Valdivia, S. A. P. Análisis temporal y espacial de la calidad del aire determinado por material particulado pm10 y pm2,5 en lima metropolitana. *Anales Científicos* **77**, 273–283, DOI: http://dx.doi.org/10.21704/ac.v77i2.699 (2016).
4. Ordóñez-Aquino, C. & Sánchez-Ccoyllo, O. Caracterización química-morfológica del pm2, 5 en lima metropolitana mediante microscopía electrónica de barrido (meb). *Acta Nova* **8**, 397–420 (2018).
5. Organization, W. H. *Air quality guidelines: global update 2005: particulate matter, ozone, nitrogen dioxide, and sulfur dioxide* (World Health Organization, 2006).
6. Vahlsing, C. & Smith, K. R. Global review of national ambient air quality standards for pm 10 and so 2 (24 h). *Air Qual. Atmosphere & Heal.* **5**, 393–399, DOI: https://doi.org/10.1007/s11869-010-0131-2 (2012).
7. SINIA. Reglamento de estandares nacionales de calidad ambiental del aire. Tech. Rep., MINAM (2001).
8. EPA-US. National ambient air quality standards for particulate matter. Tech. Rep. 10, EPA (2013).
9. Chen, Y., Shi, R., Shu, S. & Gao, W. Ensemble and enhanced pm10 concentration forecast model based on stepwise regression and wavelet analysis. *Atmospheric Environ.* **74**, 346–359, DOI: https://doi.org/10.1016/j.atmosenv.2013.04.002 (2013).
10. Kim, Y., Fu, J. S. & Miller, T. L. Improving ozone modeling in complex terrain at a fine grid resolution: Part i–examination of analysis nudging and all pbl schemes associated with lsms in meteorological model. *Atmospheric Environ.* **44**, 523–532 (2010).
11. Chen, J. *et al.* Seasonal modeling of pm2. 5 in california's san joaquin valley. *Atmospheric environment* **92**, 182–190, DOI: https://doi.org/10.1016/j.atmosenv.2014.04.030 (2014).
12. Saide, P. E. *et al.* Forecasting urban pm10 and pm2. 5 pollution episodes in very stable nocturnal conditions and complex terrain using wrf–chem co tracer model. *Atmospheric Environ.* **45**, 2769–2780, DOI: https://doi.org/10.1016/j.atmosenv.2011.02.001 (2011).
13. Li, X., Peng, L., Hu, Y., Shao, J. & Chi, T. Deep learning architecture for air quality predictions. *Environ. Sci. Pollut. Res.* **23**, 22408–22417, DOI: https://doi.org/10.1007/s11356-016-7812-9 (2016).
14. Li, C., Hsu, N. C. & Tsay, S.-C. A study on the potential applications of satellite data in air quality monitoring and forecasting. *Atmospheric Environ.* **45**, 3663–3675, DOI: https://doi.org/10.1016/j.atmosenv.2011.04.032 (2011).
15. Guarnaccia, C. *et al.* Arima models application to air pollution data in monterrey, mexico. *AIP Conf. Proc.* **1982**, 020041, DOI: 10.1063/1.5045447 (2018).
16. Adams, M. D. & Kanaroglou, P. S. Mapping real-time air pollution health risk for environmental management: Combining mobile and stationary air pollution monitoring with neural network models. *J. Environ. Manag.* **168**, 133–141, DOI: https://doi.org/10.1016/j.jenvman.2015.12.012 (2016).
17. Croitoru, C. & Nastase, I. A state of the art regarding urban air quality prediction models. *E3S Web Conf.* **32**, 01010, DOI: https://doi.org/10.1016/j.jenvman.2015.12.012 (2018).
18. Salini Calderón, G. & Pérez Jara, P. Estudio de series temporales de contaminación ambiental mediante técnicas de redes neuronales artificiales. *Ingeniare. Revista chilena de ingeniería* **14**, 284–290 (2006).
19. Guzmán, A. A. E. *et al.* Artificial neural network modeling of PM10 and PM2.5 in a tropical climate region: San Francisco de Campeche, Mexico. *Quimica Nova* **40**, 1025–1034, DOI: http://doi.org/10.21577/0100-4042.20170115 (2017).
20. Kök, İ., Şimşek, M. U. & Özdemir, S. A deep learning model for air quality prediction in smart cities. In *2017 IEEE International Conference on Big Data (Big Data)*, 1983–1990 (IEEE, 2017).
21. Jacinto Herrera, R. T. *Redes neuronales para predicción de contaminación del aire en Carabayllo-Lima*. Master's thesis, Universidad Nacional Federico Villarreal (2019).
22. Athira, V., Geetha, P., Vinayakumar, R. & Soman, K. Deepairnet: Applying recurrent networks for air quality prediction. *Procedia Comput. Sci.* **132**, 1394–1403, DOI: https://doi.org/10.1016/j.procs.2018.05.068 (2018). International Conference on Computational Intelligence and Data Science.

23. Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. The kdd process for extracting useful knowledge from volumes of data. *Commun. ACM* **39**, 27–34, DOI: 10.1145/240455.240464 (1996).

24. Rojas, C. S. A. *Condições meteorológicas e níveis de poluição na Região Metropolitana de Lima–Perú*. Master's thesis, Universidad de Sao Paulo (2013).

25. INEI. Instituto nacional de estadistica e informatica (2020).

26. Sánchez Ccoyllo, O. & Ordoñez Aquino, C. Evaluación de la calidad del aire en lima metropolitana 2015. Tech. Rep., Dirección de Meteorología y Evaluación Ambiental Atmosférica-SENAMHI (2016). Accessed on 19-07-2021.

27. Silva, J. *et al.* Particulate matter levels in a south american megacity: the metropolitan area of lima-callao, peru. *Environ. monitoring assessment* **189**, 635, DOI: https://doi.org/10.1007/s10661-017-6327-2 (2017).

28. Navares, R. & Aznarte, J. L. Predicting air quality with deep learning lstm: Towards comprehensive models. *Ecol. Informatics* **55**, 101019, DOI: https://doi.org/10.1016/j.ecoinf.2019.101019 (2020).

29. Rivera Poma, J. M. Desarrollo de un modelo dinámico para determinar la incidencia de los factores contaminantes del aire en la población de lima metropolitana. *Ind. Data* **15**, 054–062, DOI: 10.15381/idata.v15i2.6372 (2012).

30. Encalada-Malca, A. A., Cochachi-Bustamante, J. D., Rodrigues, P. C., Salas, R. & López-Gonzales, J. L. A spatio-temporal visualization approach of pm10 concentration data in metropolitan lima. *Atmosphere* **12**, 609, DOI: 10.3390/atmos12050609 (2021).

31. Royston, P. & White, I. R. Multiple imputation by chained equations (mice): Implementation in stata. *J. Stat. Softw.* **45**, 1–20, DOI: 10.18637/jss.v045.i04 (2011).

32. Seabold, S. & Perktold, J. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, vol. 57, 61 (Austin, TX, 2010).

33. Allende, H., Moraga, C. & Salas, R. Artificial neural networks in time series forecasting: a comparative analysis. *Kybernetika* **38**, 685–707 (2002).

34. Allende, H., Salas, R., Torres, R. & Moraga, C. Modular neural network applied to non-stationary time series. In Reusch, B. (ed.) *Computational Intelligence, Theory and Applications, International Conference 8th Fuzzy Days, Dortmund, Germany, Sept. 29 - Oct. 01, 2004*, vol. 33 of *Advances in Soft Computing*, 585–598, DOI: 10.1007/3-540-31182-3_54 (Springer, 2004).

35. Veloz, A., Salas, R., Allende-Cid, H. & Allende, H. Sifar: Self-identification of lags of an autoregressive tsk-based model. In *2012 IEEE 42nd International Symposium on Multiple-Valued Logic*, 226–231 (IEEE, 2012).

36. Vivas, E., Allende-Cid, H., Salas, R. & Bravo, L. Polynomial and wavelet-type transfer function models to improve fisheries' landing forecasting with exogenous variables. *Entropy* **21**, 1082, DOI: 10.3390/e21111082 (2019).

37. Vivas, E., Allende-Cid, H. & Salas, R. A systematic review of statistical and machine learning methods for electrical power forecasting with reported mape score. *Entropy* **22**, 1412, DOI: 10.3390/e22121412 (2020).

38. Morales, Y., Querales, M., Rosas, H., Allende-Cid, H. & Salas, R. A self-identification neuro-fuzzy inference framework for modeling rainfall-runoff in a chilean watershed. *J. Hydrol.* **594**, 125910, DOI: https://doi.org/10.1016/j.jhydrol.2020.125910 (2021).

39. Xayasouk, T., Lee, H. & Lee, G. Air pollution prediction using long short-term memory (lstm) and deep autoencoder (dae) models. *Sustainability* **12**, 2570, DOI: 10.3390/su12062570 (2020).

40. Graves, A. Generating sequences with recurrent neural networks (2013).

41. Bengio, S., Vinyals, O., Jaitly, N. & Shazeer, N. Scheduled sampling for sequence prediction with recurrent neural networks (2015).

42. Reddy, V., Yedavalli, P., Mohanty, S. & Nakhat, U. Deep air: Forecasting air pollution in beijing, china (2018).

43. Lipton, Z. C., Berkowitz, J. & Elkan, C. A critical review of recurrent neural networks for sequence learning (2015).

44. Li, W. *et al.* Prediction of dissolved oxygen in a fishery pond based on gated recurrent unit (gru). *Inf. Process. Agric.* **8**, 185–193, DOI: https://doi.org/10.1016/j.inpa.2020.02.002 (2021).

45. Ayturan, Y. A., Ayturan, Z. C. & Altun, H. O. Air pollution modelling with deep learning: a review. *Int. J. Environ. Pollut. Environ. Model.* **1**, 58–62 (2018).

46. Yusof, N. F. F. M. *et al.* Monsoonal differences and probability distribution of pm 10 concentration. *Environ. monitoring assessment* **163**, 655–667, DOI: https://doi.org/10.1007/s10661-009-0866-0 (2010).

47. Lasheras, F. S., Nieto, P. J. G., Gonzalo, E. G., Bonavera, L. & de Cos Juez, F. J. Evolution and forecasting of pm10 concentration at the port of gijon (spain). *Sci. Reports* **10**, 1–12, DOI: https://doi.org/10.1038/s41598-020-68636-5 (2020).

48. Delgado-Villanueva, A. & Aguirre-Loayza, A. Modelamiento y evaluación del nivel de calidad del aire mediante el análisis de grey clustering, estudio de caso lima metropolitana. *Tecnia* **30**, 114–120, DOI: http://dx.doi.org/10.21754/tecnia.v30i1.588 (2020).

49. Silva, J. S., Rojas, J. P., Norabuena, M. & Seguel, R. J. Ozone and volatile organic compounds in the metropolitan area of lima-callao, peru. *Air Qual. Atmosphere & Heal.* **11**, 993–1008, DOI: https://doi.org/10.1007/s11869-018-0604-2 (2018).

50. Moreno Jiménez, A., Méndez Arranz, D. *et al.* La concentración de partículas en el aire: análisis estadístico de la relación espacial entre medidas de superficie y del sensor modis para dos tipos de tiempo en la comunidad de madrid. *Investig. Geográficas* **73**, 189–209, DOI: https://doi.org/10.14198/INGEO2020.MJCTMA (2020).

51. Sahin, F., Kara, M. K., Koc, A. & Sahin, G. Multi-criteria decision-making using gis-ahp for air pollution problem in igdir province/turkey. *Environ. Sci. Pollut. Res.* **27**, 36215–36230, DOI: https://doi.org/10.1007/s11356-020-09710-3 (2020).

52. Taheri Shahraiyni, H. & Sodoudi, S. Statistical modeling approaches for pm10 prediction in urban areas; a review of 21st-century studies. *Atmosphere* **7**, 15, DOI: 10.3390/atmos7020015 (2016).

53. Ilizarbe-Gonzáles, G. M. *et al.* Chemical characteristics and identification of pm10 sources in two districts of lima, peru. *Dyna* **87**, 57–65, DOI: https://doi.org/10.15446/dyna.v87n215.83688 (2020).

54. OTD. Tránsito de vehículos a nivel nacional aumentó 15,5%. Tech. Rep. 076, INEI (2018).

55. Jain, S., Sharma, S., Vijayan, N. & Mandal, T. Seasonal characteristics of aerosols (pm2.5 and pm10) and their source apportionment using pmf: A four year study over delhi, india. *Environ. Pollut.* **262**, 114337, DOI: https://doi.org/10.1016/j.envpol.2020.114337 (2020).

56. Alolayan, M. A., Brown, K. W., Evans, J. S., Bouhamra, W. S. & Koutrakis, P. Source apportionment of fine particles in kuwait city. *Sci. total environment* **448**, 14–25, DOI: https://doi.org/10.1016/j.scitotenv.2012.11.090 (2013).

57. Owoade, K. O. *et al.* Chemical compositions and source identification of particulate matter (pm2. 5 and pm2. 5–10) from a scrap iron and steel smelting industry along the ife–ibadan highway, nigeria. *Atmospheric Pollut. Res.* **6**, 107–119, DOI: https://doi.org/10.5094/APR.2015.013 (2015).

58. Capel Molina, J. J. Lima, un clima de desierto litoral. *Anales de Geografía de la Universidad Complutense* **19**, 25 (1999).

59. Concepción, E. & Rodríguez, J. Informe nacional de la calidad del aire 2013-2014. https://bit.ly/36KTRAM/ (2014). Accessed on 18-07-2021.

60. Iparraguirre Medina, R. L. & Valdivia Torres, A. G. *Caracterización y problemática de las ladrilleras en Huachipa-Lurigancho-Lima. 2018*. Master's thesis, Universidad Católica Sedes Sapientiae, http://repositorio.ucss.edu.pe/handle/UCSS/735 (2018). Accessed on 18-07-2021.

61. Álvarez-Liébana, J. & Ruiz-Medina, M. Prediction of air pollutants pm 10 by arbx (1) processes. *Stoch. Environ. Res. Risk Assess.* **33**, 1721–1736, DOI: https://doi.org/10.1007/s00477-019-01712-z (2019).

62. Mellado, D., Saavedra, C., Chabert, S., Torres, R. & Salas, R. Self-improving generative artificial neural network for pseudorehearsal incremental class learning. *Algorithms* **12**, DOI: 10.3390/a12100206 (2019).

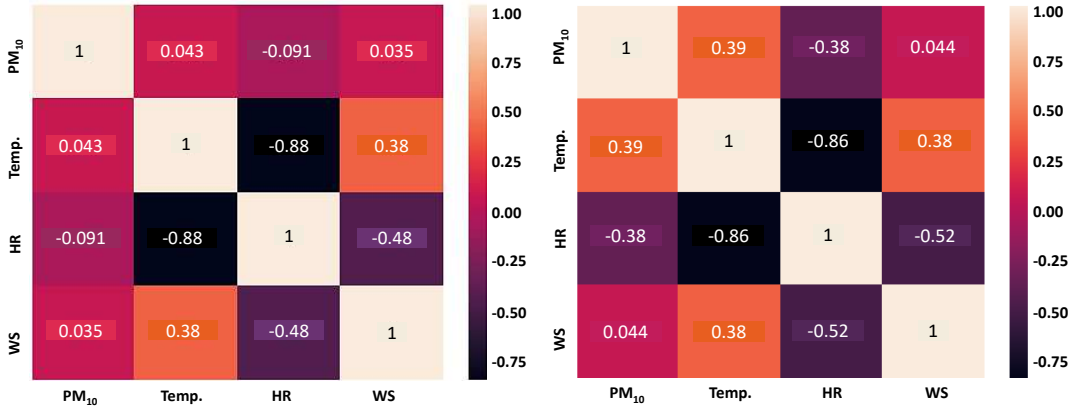
63. Veloz, A., Salas, R., Allende-Cid, H., Allende, H. & Moraga, C. Identification of lags in nonlinear autoregressive time series using a flexible fuzzy model. *Neural Process. Lett.* **43**, 641–666, DOI: https://doi.org/10.1007/s11063-015-9438-1 (2016).

Acknowledgements

Javier Linkolk López-Gonzales acknowledges financial support from the ANID scholarship. The work of Chardin Hoyos Cordova and Manuel Niño Lopez Portocarrero were supported by the grant *Beca 18* of the national government. P.C. Rodrigues acknowledges financial support from the Brazilian National Council for Scientific and Technological (CNPq) grant "bolsa de produtividade PQ-2" 305852/2019-1. The authors are grateful to the *Servicio Nacional de Meteorología e Hidrología* (SENAMHI) for providing the air quality data used in this study.

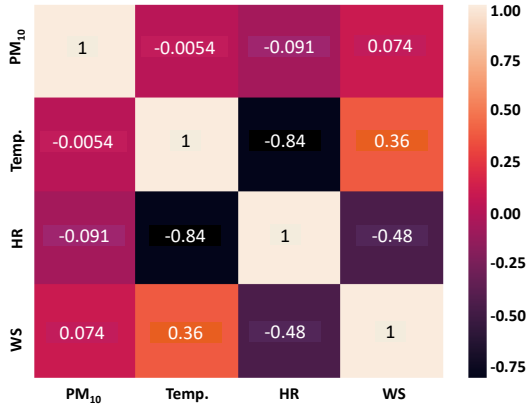
Author contributions statement

J.L.L-G., C.H.C. and M.N.L.P conceived the experiment(s), R.S., R.T. and P.C.R. conducted the experiment(s), C.H.C., R.S., J.L.L-G and P.C.R. analysed the results. All authors reviewed the manuscript.

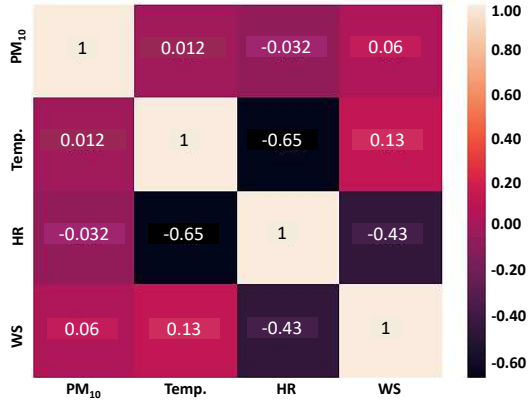


(a) ATE

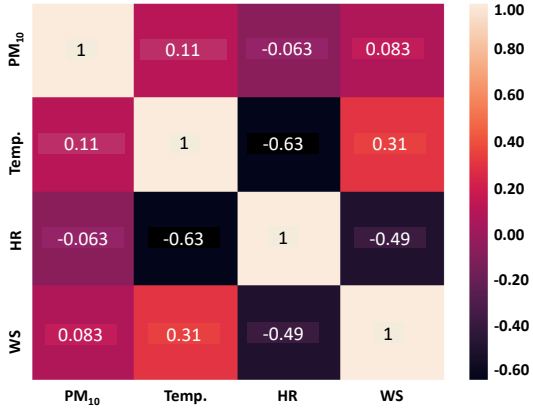
(b) HCH



(c) CRB



(d) CDM



(e) SMP

Figure 7. Correlation matrices between the meteorological variables and the PM_{10} for each monitoring station.

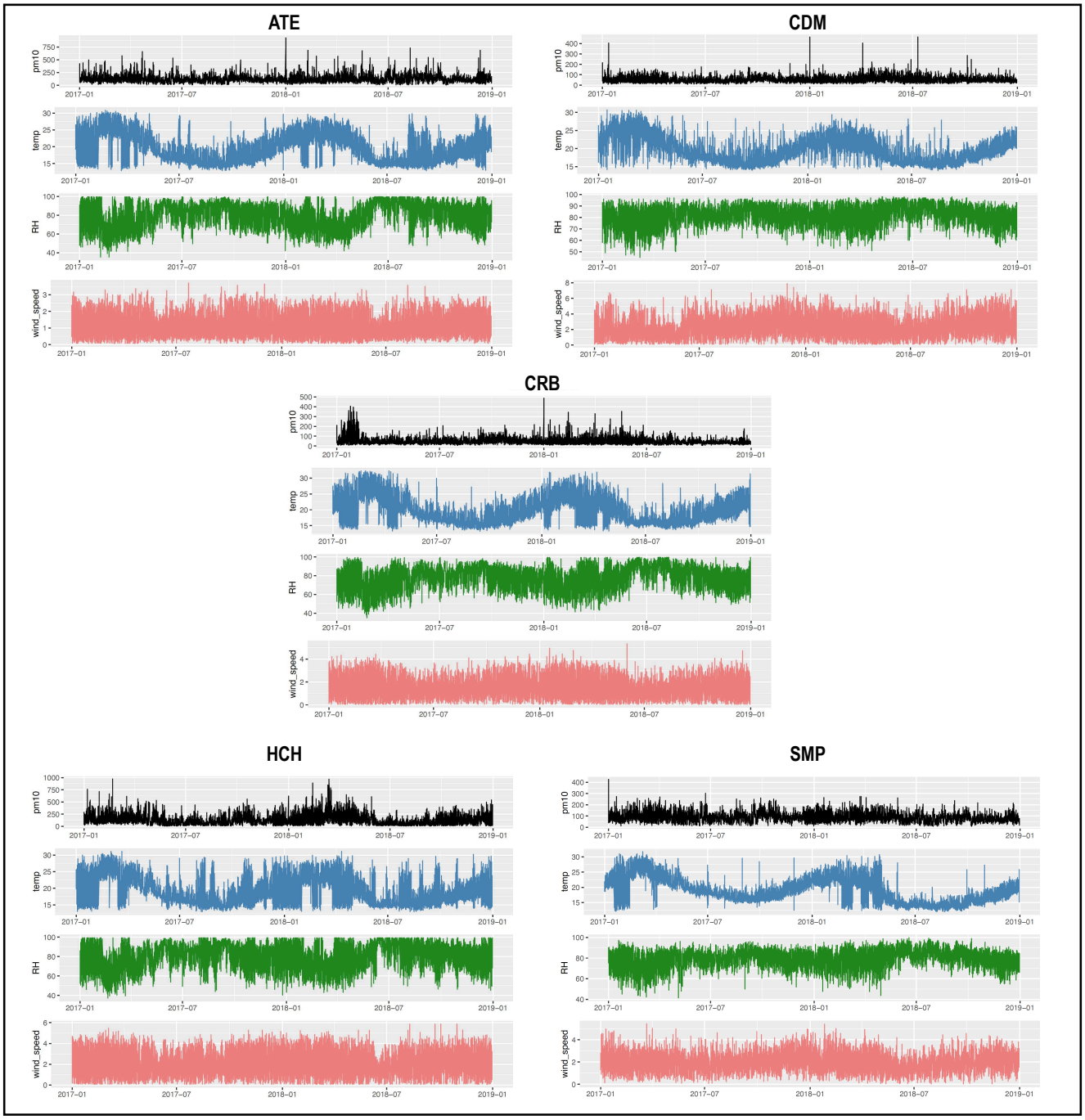


Figure 8. Time series of all variables, PM_{10}, temperature, relative humidity and wind speed, in each monitoring station, ATE, CDM, CRB, HCH and ATE, respectively.

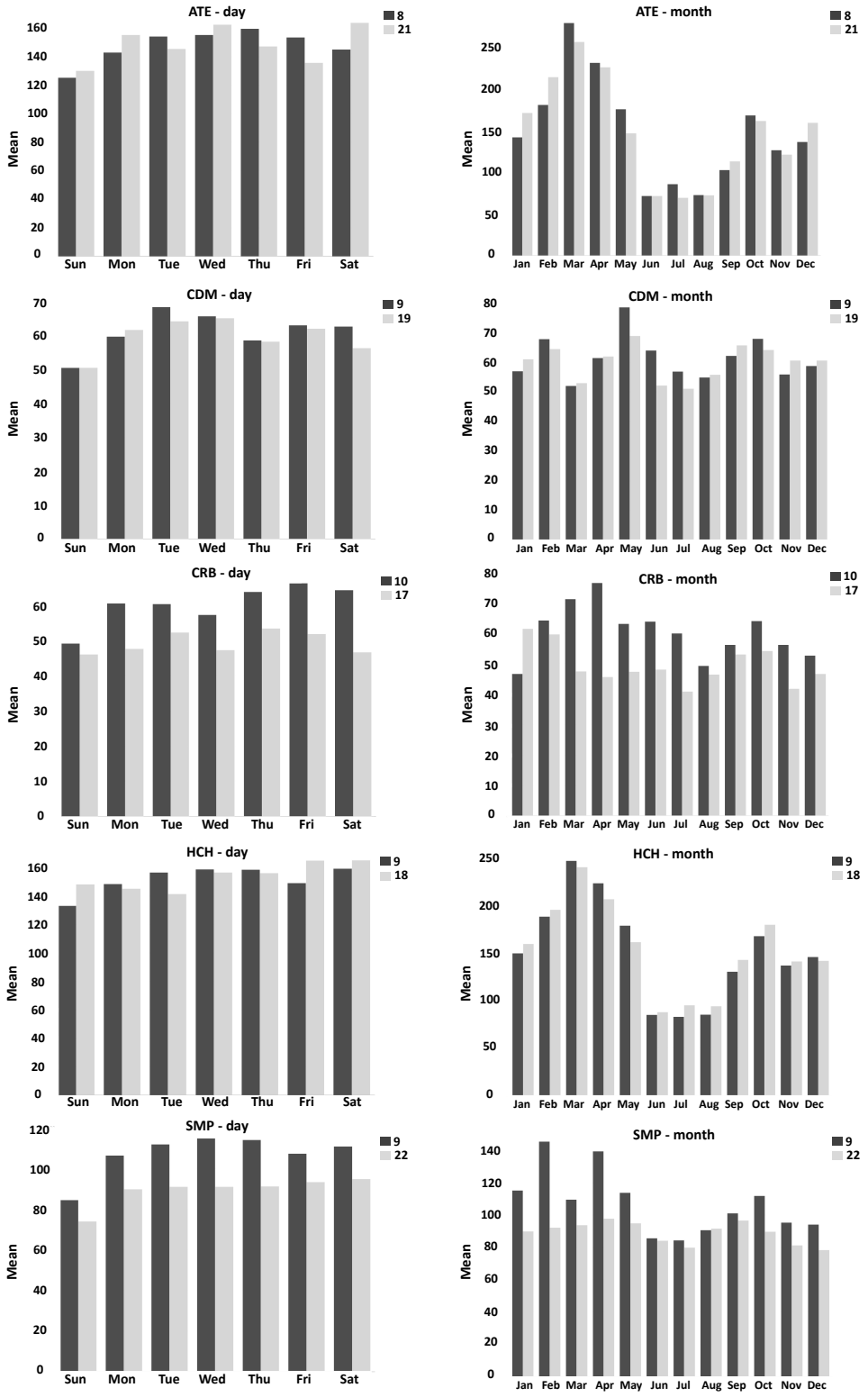


Figure 9. Bar plot per day and per month for each monitoring station, ATE, CDM, CRB, HCH and ATE, respectively. The average hourly pollution per day of the week, and per month of the year, is reported for all stations.

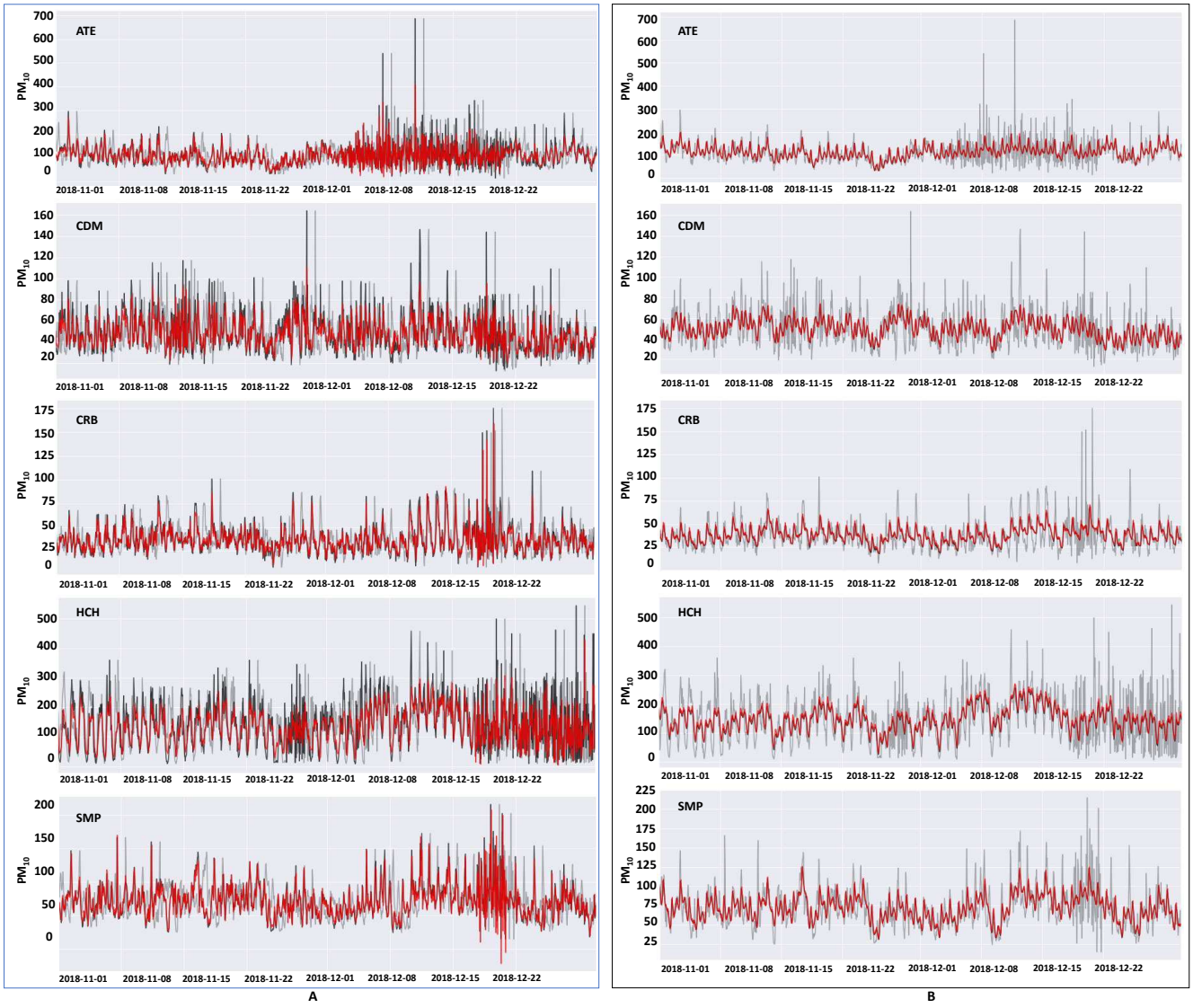


Figure 10. Forecast values with the LSTM for all considered stations, **(A)** no decomposition and **(B)** with decomposition. The LSTM network expects the input data (X) to be provided with a specific matrix structure in the form of: samples, time stages and characteristics. The sigmoid activation function was used for the LSTM memory blocks.